Change Our View: A Distributed Topic Modelling and Feature-Based Study of Persuasion Dynamics in `r/ChangeMyView`

39375, 41830, 44051, 42309

KEYWORDS

Distributed computing, Topic modelling, Persuasion, Reddit, Opinion change, PySpark, NLP Pipelines, LDA, r/ChangeMyView (CMV)

1 ABSTRACT

This study probes the mechanics of opinion change in the *r/ChangeMyView* subreddit (CMV) by combining large-scale text mining with distributed machine-learning pipelines. Using the Webis-CMV-20 corpus including 65,169 discussion threads, we construct a PySpark workflow that cleans and merges posts metadata, uncovers latent topics via Latent Dirichlet Allocation (LDA), and by engineering structural, linguistic, sentiment, and temporal features, we train both single-core and distributed decision-tree ensembles that aim to predict changes in opinion. Descriptive analyses reveal how post length, interaction volume, and readability signal a higher likelihood of concession, rather than the topic or sentiment behind the post. Our best distributed model attains an F_1 score of 0.64 while scaling linearly across a 12-node Spark cluster. The distributed approach also reduces the random-forest training time from 49s to under 7s. Overall, the project shows that community interaction patterns are strongly correlated with persuasion, rather than thematic content, and provides a fully reproducible, cloud-based pipeline for future work on large social-media databases.

2 INTRODUCTION

Why do some arguments persuade while others only reinforce preexisting beliefs? Persuasion is a foundational concept in political science, marketing, and behavioural economics, and yet the study of persuasion has traditionally been limited by small-scale experiments or qualitative analyses. Further, obtaining data of faithfully occurring debates is a challenge in and of itself. As more and more discourse is happening in purpose-built online forums, Reddit's `r/ChangeMyView` (CMV) has become one of the best instances of naturally occurring and largely faithful online discourse between people of varying viewpoints. It currently has 3.8 million members and ranks among the 1% largest subreddit communities, with up to 200 posts per day. CMV presents a unique opportunity as it is explicitly designed for users to post opinions they are open to changing, and to acknowledge when their views have been altered. The wide range of topics covered in CMV makes persuasion research within this subreddit community particularly interesting. Moreover, Reddit's anonymity encourages users to be more truthful and fosters increased engagement, even during disagreements or failed debates.

Users of CMV voluntarily start a discussion by posting a view or opinion on any topic and invite community members to change their minds. The portion of CMV crucial to the study of these interactions is the assigning of a Δ badge, by the Original Poster (OP), to the/any reply that changes their mind. Our goal in leveraging such well structured data of human discourse is two-fold: (i) quantify what drives an OP to concede, and (ii) predict concession using the way in which the individual approaches a forum of online discussion, while keeping the pipeline fully distributed.

We contribute:

- a cleaned CMV corpus of $65,169$ original posts;
- a 45-topic Spark LDA fitted to the entire corpus and manually mapped to 14 semantic categories;
- an interpretable feature set that describes *what* and *how* OPs approach the forum of discussion with the potential to have their mind changed;
- additional original features such as sentiment analysis and time series comparisons to see how online forums reflect key world events (elections, wars, sports);
- a comparison between single-core and multi-core decision trees, showing that distribution matters more than depth in this imbalanced setting;

Our work leverages distributed computing techniques to analyse the CMV corpus, employs topic modelling to understand the thematic landscape of discussions, and derives linguistic features that may signal increased likelihood of OP persuasion. The distributed nature of our approach allows us to process this large dataset efficiently while maintaining the ability to capture complex patterns in persuasive discourse and draw insights from them.

3 RELATED WORK

3.1 Online Opinion Shaping

Previous work on online persuasion has explored various platforms and contexts. Tan et al. [11] conducted a large-scale study on `r/ChangeMyView`, identifying linguistic features associated with persuasive arguments on a data set from 2013 to 2015. Their work showed that successful arguments tended to be longer (in terms of length of comment), used different vocabulary from the original post, and employed specific rhetorical strategies. A large part of the research was focused on both content within the original posts and the comments. The paper was innovative in two ways. First, it paired OPs with their successful argument, meaning that features could be extracted for each angle and provide comprehensive interaction dynamics. The authors found that the first two challengers were 3 times more likely to successfully convince the original author compared to the tenth. Second, they provided in-depth information about the challengers, showing that more experienced challengers (active community members) were more successful at convincing than others. The paper serves as the backbone for our own research. Lukin et al. [6] extended work on online persuasion by examining the role of argumentative structure and personality traits in persuasion effectiveness. These approaches did not incorporate any topic modelling, limiting their ability to uncover potentially different dynamics within certain subjects. Additionally, a Pew Research Center's 2016 report brings evidence that Reddit functions as a significant hub for news and information [7]. CMV's structure provides an opportunity for users to both engage in current debates and inform themselves, a valuable source of public opinion and societal trends neglected by the mentioned papers.

Furthermore, the study conducted by Ott and Aoki [9] investigates how different media platforms influence opinion change, focusing on the persuasive power of tweets versus traditional news articles. The researchers found that tweets expressing moral outrage tend to receive more engagement, which in turn reinforces the use of emotional appeals in persuasive communication. Moreover, the perceived credibility of the message source significantly affects its persuasive impact, with news articles being rated as more persuasive and trustworthy than tweets, even when presenting identical arguments. These findings are directly relevant to our project, as they highlight the interplay between emotional tone, source credibility, and medium effects in shaping persuasive success—dimensions that complement our analysis of structured debates on Reddit's r/ChangeMyView.

3.2 Topic Modelling in Social Media

Topic modelling techniques have been widely applied to social media data to discover latent themes. Blei et al. [1] introduced Latent Dirichlet Allocation (LDA), which has since become a standard approach for unsupervised topic discovery. Additionally, practical implementations of large-scale topic modelling with PySpark and Spark NLP have been demonstrated in applied engineering settings [8], underscoring the feasibility of scalable pipelines for high-volume social-media analytics.

3.3 Distributed Computing for NLP

The application of distributed computing frameworks to natural language processing tasks has gained traction with the increasing volume of social media data. Zaharia et al. [13] demonstrated how Apache Spark can be used for large-scale processing tasks. Wang et al. [12] specifically explored the use of distributed LDA implementations for topic modelling of social media conversations. Our work builds on these approaches by creating an end-to-end distributed pipeline for analyzing persuasion dynamics.

Additionally, we draw inspiration from innovative examples such as those presented in Isaac Triguero's textbook[4], where Spark is applied to build both local and global models using Resilient Distributed Datasets (RDD) and MLlib APIs, respectively. In particular, the progression from a baseline scikit-learn model to distributed versions illustrates how Neuro-linguistic programming (NLP) and classification workflows can scale using Spark's architecture. This principle guided the structure of our own distributed system, especially in deploying decision tree classifiers and feature engineering at scale across our CMV corpus.

3.4 Predicting Opinion Change

Predicting whether an argument will change someone's opinion remains challenging. Hidey and McKeown [2] developed models to predict successful persuasion, focusing on modelling the sequence of arguments in social media posts using neural models with embeddings for words, discourse relations, and semantic frames.

Correlation vs. causation. Finally, disentangling causal effects from observational traces generally requires either controlled interventions or strong identification assumptions [3, 10]. Like Tan *et al.*, we restrict our claims to correlation of persuasion rather than causation. Nonetheless, our distributed architecture is a prerequisite for future causal experiments. For instance, randomizing reply exposure or order at scale—thereby paving the way from "patterns that co-occur" to "mechanisms that make a difference".

4 METHODOLOGY

We ensured our project was reproducible and interpretable end-to-end across five structured notebooks, all designed to run within the LSE Google Cloud Platform (GCP) environment. To fully understand the workflow, we recommend reviewing the project pipeline in Appendix Figure 2.

4.1 Data Acquisition & Pre-processing

Reddit severely limits the use of its API to 1000 observations with very limited scraping opportunity. We therefore use the Webis-CMV 20 dataset [5], a JSON dump that contains 65 169 posts from the subreddit `r/ChangeMyView`; from January 2013 to September 2017 uploaded on a shared bucket on GCP. All notebooks' output artefacts were also saved to the same shared bucket. The file `threads.jsonl` stores each discussion thread as an independent JSON object enriched with relevant post metadata fields summarized in Table 1.

Table 1: Schema excerpt for `threads.jsonl`.

Field	Description
`title`	Post headline (argument synopsis)
`selftext`	Main body of the original post
`num_comments`	Total number of comments in the thread
`score`	Reddit score (up-votes minus down-votes)
`delta`	`true` if OP granted a Δ badge
`urls[]`	Array of outbound links mentioned in the post
`name`	Unique Reddit ID of the author
`created_utc`	Post timestamp

Cleaning pipeline: Raw data are staged on Google Cloud Storage and ingested by a 12 × `n2-standard-4` Dataproc cluster running Spark 3.5. We execute a shell script `my_actions.sh` to install required dependencies, including the `nltk` library and in particular the VADER lexicon, essential for performing sentiment analysis in subsequent stages. An *explicit* `StructType`, mirroring Table 1, which prevents latency and type mismatches introduced by schema inference.
To produce a fully cleaned, tokenised, and timestamp-aligned dataset, suitable for scalable downstream analysis in Spark, we performed these transformations in a distributed fashion using PySpark and some PySpark SQL functions:

(1) **Timestamp normalisation:** Convert the `created_utc` epoch field to Spark-native `timestamp` type, and extract a `year_month` string to support temporal analysis.

(2) **Moderator filter:** Remove in-line moderator messages, unrelated to the post.
(3) **Field merge:** Concatenate `title` and `selftext` to grasp the entire author's post.
(4) **General text preprocessing:** Lowercase the input string and and strip punctuation, digits, and non-letter characters.
(5) **Custom stop-words:** In addition to NLTK's pre-made english stopwords function, a specific list of words were removed through trial and error due to high repetition on r/CMV (*change, view, opinion*, ...).
(6) **Tokenisation & lemmatisation** split into tokens and lemmatise each token using `nltk`.
(7) **Join tokens:** The surviving tokens are rejoined into a space-separated string stored in `processed`.

The resulting DataFrame (62,843 OPs × 12 columns) was successfully pre-processed in order to continue the Topic Discovery covered in the next section.

4.2 Topic Modelling

To identify the high-level themes present in the corpus, we implemented a distributed Latent Dirichlet Allocation (LDA) model using Spark MLlib.

We encode each token array as a sparse bag-of-words vector $\mathbf{x}_d$ using Spark's `CountVectorizer`. The vocabulary size V is controlled by a `minDF` (minimum document frequency) threshold `minDF` = 8. Terms occurring in fewer than eight posts are discarded because (i) rare tokens are often typos, inflected slang, or idiosyncratic user names that add noise; (ii) removing them reduces the feature space significantly (iii) it improves topic coherence by forcing the model to rely on words that generalise across multiple discussions.

Latent Dirichlet Allocation: On the resulting document–term matrix we train a probabilistic topic model with ($K = 45$) latent topics. Pilot experiments with $K \in \{20, 30, 40, 45, 50\}$ showed that < 45 forces semantically distinct debates (e.g. climate versus energy policy) to merge, whereas > 45 splits cohesive themes into nearly duplicate sub-topics. Our selection offers the best compromise between granularity and interpretability.

LDA returns for every post d:

- a length-K probability vector $\boldsymbol{\gamma}_d$, whose entries sum to 1,
- the dominant topic $k^* = \arg\max_k \gamma_{d,k}$.

For each topic we inspected its eight highest-probability words. Topics sharing a coherent semantic field were grouped into *macro-categories* such as *Politics*, *Gender*, or *Religion*; genuinely ambiguous clusters are tagged *Other*. The label set enables downstream analyses without revisiting raw token lists.

Output artefacts. The augmented DataFrame then contained:

`features` sparse term-frequency vector (input to LDA);
`topicDistribution` full $\boldsymbol{\gamma}_d$;
`dominant_topic` index k^*;
`category_title` human-readable macro-category.

These artefacts were saved in several partitions to Google Cloud Storage and fed the feature-engineering and modelling notebooks that follow.

4.3 Feature Engineering

In order to complement our topic-based representation of CMV posts, we constructed a set of interpretable features designed to capture how authors express themselves, how structured or emotional their posts are, and how often others engage with them.

Each feature was chosen to reflect a plausible mechanism behind persuasion such as confidence, clarity, emotional tone, or evidential support. All transformations were applied in a fully distributed Spark environment using UDFs and vectorized operations, ensuring scalability across tens of thousands of posts.

The engineered features used in downstream prediction models are:

Length and structure:

- **Post length:** — Total number of whitespace-separated tokens in the main post body (`selftext`). Serves as a proxy for argument depth, verbosity or conviction.
- **Title length:** Token count in the headline (`title`). Longer titles may signal specificity or emotionally loaded messages.
- **URL presence:** A binary flag indicating whether the post contains one or more hyperlinks. Used as a coarse proxy for the inclusion of external evidence.

Linguistic and stylistic markers:

- **Flesch–Kincaid Grade (`fk_grade`)** — A readability score computed from sentence and syllable structure:

  $$\text{FK Grade} = 0.39 \times \text{ASL} + 11.8 \times \text{ASW} - 15.59$$

  where ASL is average sentence length and ASW is average syllables per word. Most scores range from approximately ~ 6 (simple) to > 12 (college-level), potentially reflecting the cognitive effort of processing an argument. The function is unbounded, allowing for out-of-range values, which we leverage to identify and filter out invalid posts. Topics often related to science or healthcare scored higher Flesch-Kincaid grades than sports.
- **First-person singular count** — Frequency of self-referential terms such as "I", "me", or "my". These often convey introspection or personal stakes.
- **First-person plural count** — Frequency of collective pronouns like "we" or "our", which suggest group alignment or shared responsibility.

Sentiment:

- **Compound sentiment score (`sentiment`)** — Extracted using VADER, this scalar captures the emotional tone of the post, ranging from −1 (very negative) to +1 (very positive). We selected VADER over TextBlob due to its specialization in social media content, which aligned more closely with the nature of our Reddit data. Although our preprocessing pipeline removed punctuation for standardization, VADER is designed to catch features like exclamation marks and emojis that are common in Reddit to enhance sentiment detection. Posts with neutral or calm sentiment considered between -0.5 and 0.5 are processed by the absence of emotionally charged features.

Engagement and feedback.

- **Score (score)** — Reddit score (upvotes minus downvotes), as a rough measure of crowd approval.
- **Number of comments (num_comments)** — Volume of engagement a post receives. This includes back and forths between the OP author and other users. May correlate with controversy or clarity.

These features allowed us to model not only what an author says, but how they say it, how often others respond, and under what conditions persuasion is likely to occur. All variables are stored in a single denormalized table and cached for downstream modelling.

4.4 Predictive Models

We trained several distributed and standard machine learning models to predict whether a comment would receive a delta:

- Logistic Regression
- Random Forest
- Distributed Decision Trees

These models varied in complexity and scalability, which allowed us to evaluate both baseline and more robust ensemble-based classification techniques under the constraints of class imbalance and interpretability.

Modelling Approach:

To predict whether a comment received a *delta*, we implemented supervised classification models such as decision trees and random forests. The target variable was binary, indicating the presence or absence of a delta. The dataset exhibited notable class imbalance, with the positive class (delta) representing approximately 14% of the observations.

The baseline models trained include:

- **Decision Tree Classifier** using a maximum depth of 10 to prevent overfitting.
- **Random Forest Classifier** composed of 500 estimators, each with a maximum depth of 10.

To address class imbalance, we evaluated models both with and without the use of class weights, which automatically adjusted weights inversely proportional to class frequencies.

The dataset was split into training and testing subsets using a 70/30 ratio, with a fixed random seed to ensure reproducibility. Model performance was assessed using standard classification metrics: precision, recall, F1-score, and accuracy. In particular, we emphasized the weighted F1-score due to its ability to balance performance across imbalanced classes.

4.5 Distributed Tree-Based Learning

To evaluate model scalability and test the effectiveness of ensemble-like parallelism, we extended our tree-based models using distributed Decision Trees implemented in PySpark. The training data was repartitioned into 12 subsets. Each partition trained an independent decision tree using identical hyper-parameters as the baseline approach. During inference, predictions from each model were aggregated using majority voting.

This local model parallelism mimics traditional bagging methods by increasing diversity across models without requiring full bootstrap sampling. Models were trained using the scikit-learn API inside each partition, and final model lists were broadcast to all worker nodes for efficient prediction.

4.6 Exploratory Analysis

We explored patterns in our data and extracted features through aggregate tables and visualizations such as time series plots to examine relationships between topics, linguistic features, and persuasion outcomes.

Delta rate and number of comments: We produced a set of aggregate tables, of which two were particularly insightful. Using PySpark SQL, Figure 5 shows comment counts and delta rate prevalence between categories. The plot indicates a correlation between the amount of comments and the likelihood of the OP to assign a delta. This relationship was expected since the greater the number of opposing opinions, the more often the OP could be convinced.

Sentiment score by category: Certain categories, such as politics and society, tended to generate more comments, suggesting higher engagement or potentially polarizing behaviour. Going beyond the volume of comments, we examined sentiment. Online users often adapted their language according to the topic, with discussions about politics being more emotionally charged than those about food or animals. We therefore suspected that our sentiment score feature would differ accordingly across categories. Here, this score ranges from -1 to 1, where negative values indicate more negative sentiment and positive values positive sentiment. Our boxplots (Figure 3) revealed that while some categories like food and society skewed more positive, others like politics were skewed slightly more negative, the overall differences in sentiment across categories were modest.

Sentiment score by category over time: Given that OPs often reflected current-day events, we decided to look into whether OP sentiment would change across time according to certain categories. For instance figure [7] shows the ratio between number of highly positive posts and number of highly negative posts between 2016 and 2017 for three categories: politics, gender, society. We focused on the chosen period, due to the 2016 United States' Presidential Election[1]. What we can observe from this plot aligns with the intuition that the 2016 US election caused divisiveness, as the ratio of negative to positive sentiment intensely swings from November to December 2016 for the category *Gender*.

5 NUMERICAL RESULTS

5.1 Topic Discovery

The thematic categories used in the analysis were assigned through a manual labelling process of latent topics extracted from the data. Each topic was defined by its top contributing keywords and subsequently assigned to one of the 14 semantic categories.

For instance, the following topic was labelled as Politics based on its dominant terms:

Topic 1:

- government: 0.0110
- country: 0.0098
- state: 0.0091

[1] We assumed that President Trump's victory would have affected the volatility of sentiment for many posts.

- gun: 0.0089
- law: 0.0077
- war: 0.0066
- police: 0.0061
- military: 0.0048

The numbers next to each word are the probability of that word being generated if we sampled a single token conditioned on this specific topic. From this topic, there is then a 1.1% chance that the word generated would be "government".

The labelling approach was applied consistently across all topics to produce the final category assignments shown in Table 2. Notice that the label 'Others' is created for all topics within the 45 generated by LDA that do not seem to have an interpretable category to assign it to, whereas 'Other categories' include smaller, and very easy to identify topics, such as Health, Animals, Food, among others.

Table 2: Category Distribution of Posts

Category	Count	Percentage (%)
Society	22,749	28.19%
Other	20,176	25.00%
Politics	13,118	16.26%
Gender	3,518	4.36%
Environment	2,816	3.49%
Culture	1,605	1.99%
Other categories	1203	1.49%

5.2 Prediction

Initial experiments with both the Decision Tree and Random Forest classifiers revealed that, when trained without accounting for class imbalance, the models overwhelmingly predicted the majority class (*no delta*). For instance, the unbalanced Decision Tree correctly identified only 68 deltas out of 2,786, while the Random Forest performed even worse, with just 16 true positives. This illustrates the severity of the class imbalance (1 delta for every 6 non-deltas) and highlights the need to apply appropriate techniques—such as class weighting or balanced sampling—to ensure meaningful detection of the minority class. Therefore, all subsequent results presented in this section are based on models trained with class balancing enabled.

Weighing by class: Both models demonstrated a notable improvement in their ability to detect deltas. The trade-off came in the form of decreased precision for the majority class, with many *no delta* comments being misclassified as deltas.

In the case of the Decision Tree, recall (defined as correctly predicted number deltas, divided by the total number of actual deltas) for the minority class increased substantially from near-zero to 0.79, while precision (defined as the number of correctly predicted deltas divided by all predicted deltas by the model) remained modest at 0.23, resulting in an F1-score of 0.35 for delta classification. In other words, our model correctly identified 80% of all deltas, and was correct only 23% of the time. The Random Forest classifier showed a similar trend but with slightly better overall balance.

Table 3: Confusion Matrix Comparison: Decision Tree vs. Random Forest (with class weighting)

Actual Class	Decision Tree		Random Forest	
	No delta	Delta	No delta	Delta
No delta (0)	9308	7457	9933	6832
Delta (1)	598	2188	661	2125

It achieved an F1-score of 0.36 for deltas, with recall of 0.76 and precision of 0.24. Compared to the Decision Tree, it also retained slightly better performance on the majority class, producing a more favorable balance across precision and recall. The weighted average F1-scores improved to 0.65 for the Decision Tree and 0.67 for the Random Forest, confirming that ensemble methods can offer a modest but consistent advantage when handling imbalanced classification tasks.

Tables 4 and 5 summarize the classification results on the test set. With class weighting enabled, both classifiers improved substantially in detecting deltas.

Table 4: Decision Tree with Class Weighting

Class	Precision	Recall	F1-score
No delta (0)	0.94	0.56	0.70
Delta (1)	0.23	0.79	0.35
Macro avg	0.58	0.67	0.53
Weighted avg	0.84	0.59	**0.65**

Table 5: Random Forest with Class Weighting

Class	Precision	Recall	F1-score
No delta (0)	0.94	0.59	0.73
Delta (1)	0.24	0.76	0.36
Macro avg	0.59	0.68	0.54
Weighted avg	0.84	0.62	**0.67**

5.3 Distributed Approach

We implemented and evaluated two decision tree classifiers and two random forest classifiers using identical setups: a maximum depth of 10, balanced class weights, the same random seed, and for the random forests, 500 estimators. Class weighting was essential, as unbalanced models tended to maximize accuracy by overwhelmingly predicting the majority class (*no delta*), while failing to detect minority cases altogether.

For classifiers, we compare a baseline and a distributed architecture. The training data was partitioned into 12 subsets[2], with each partition independently training a decision tree using the same

[2] The reasoning for choosing 12 partitions was inspired by work done in class, where we used 6 partitions for 30k observations. We therefore decided to double the number of partitions

hyper-parameters. During inference, predictions from all trees were aggregated using hard majority voting. This approach mimics bagging but without bootstrap sampling, resulting in a lightweight ensemble that improves model diversity without increasing complexity.

Table 6 presents a side-by-side comparison of the decision tree confusion matrices for the baseline and distributed models. The distributed model performed slightly better in detecting deltas, with nearly identical recall and a modest reduction in false positives. These improvements suggest that ensemble-style aggregation across partitions introduced useful diversity, improving minority class detection.

Table 6: Decision Tree Confusion Matrix Comparison: Baseline vs. Distributed

Actual Class	**Baseline**		**Distributed**	
	No delta	**Delta**	**No delta**	**Delta**
No delta (0)	9308	7457	9125	7624
Delta (1)	598	2188	572	2184

Table 7: Random Forest Confusion Matrix Comparison: Baseline vs. Distributed

Actual Class	**Baseline**		**Distributed**	
	No delta	**Delta**	**No delta**	**Delta**
No delta (0)	9933	6832	13259	3490
Delta (1)	661	2125	1670	1086

Similarly, Table 7 shows the same comparison for the random forest classifier. Here, the distributed model improves in predicting true negatives at the expense of true positives. Additionally, some performance metrics are slightly worse than the baseline model: precision remained at .84, recall fell to .58 (from .62) and F1 to .64 (from .67).

5.4 Feature Importance

Understanding which features most influence the classification decision was essential to our research. Tree-based models such as Decision Trees and Random Forests provided internal feature importance scores; that reflect how frequently and effectively each feature was used to split the data across the tree or forest structure. However, it is important to note that these scores do not capture the *direction* of the feature's effect, whether increasing a feature value increases or decreases the likelihood of a delta. Logistic regression, in contrast, also provided directional coefficients, indicating whether a feature increases or decreases the probability of a delta.

These coefficients offer preliminary insight into how feature values correlate with the probability of receiving a delta. This directional understanding will guide our interpretation of user behavior and content characteristics in future sections.

Table 8 shows the top features ranked by tree-based importance, along with their direction of association from logistic regression.

Table 8: Feature Importance from Decision Tree and Random Forest, with Logistic Regression Direction

Feature	**DT Imp.**	**RF Imp.**	**Logit Coef.**
num_comments	0.592	0.492	+0.002
post_content_length	0.146	0.197	−0.011
fk_grade	0.057	0.065	−0.023
sentiment	0.048	0.058	+0.048
score	0.044	0.058	−0.009
title_content_length	0.038	0.049	+0.002
first_person_singular_count	0.028	0.042	+0.001
first_person_plural_count	0.022	0.016	+0.000

As shown, `num_comments` was by far the most influential feature in both tree-based models. Its very weak positive logistic coefficient suggested that the number of comments was not linearly associated with receiving a delta. As mentioned in previous research, the number of comments have an initial effect on receiving a delta, however that effect flattens out after 30 comments [11]. This is consistent with the idea that while a user may initially be open to changing their mind, additional comments beyond a certain point are unlikely to be persuasive, a pattern seen in the partial dependence plot in Figure 6.

Both models consistently identified the length of the post content as the second most important predictor of whether a user changed their view. While we hypothesized that the length of a post might reflect the depth or the engagement of the user, our logistic regression could not confirm it, as its coefficients were very weak.

Interestingly, the most important features were surface-level indicators of participation, which appeared to play a greater role in influencing viewpoint-change compared to linguistics (text complexity or sentiment).

Marginal Role of Categorical Dummies

To assess the predictive contribution of the topic categories, we included a set of dummy variables representing the manually labelled semantic category of each post (e.g., `Politics`, `Society`, `Gender`, etc.). These variables were encoded as one-hot indicators and entered the model alongside all other numerical features.

The results reveal that these category dummies provide minimal added value in predicting whether a comment receives a delta. In our feature importance analysis using the Random Forest model, the effects of the category indicators ranked below eighth overall—far below the leading predictors such as `num_comments` or `post_content_length`.

This suggests that topical classification, as encoded through categories, does not meaningfully inform the likelihood of a delta. Instead, delta prediction appears to rely more heavily on structural or linguistic signals, rather than the broad thematic domain of the post.

Consequently, while category dummies were retained for completeness and interpretability, they play a negligible role in the model's decision process and do not improve predictive performance in any substantial way.

Feature analysis: Grade Level

To further interpret the effect of `fk_grade` on delta predictions, we computed a partial dependence plot (PDP) using the trained decision tree model. The plot estimates the marginal effect of grade level on the predicted probability of receiving a delta, averaging over all other features.

The results, shown in Figure 1, indicate a non-monotonic relationship. The probability of receiving a delta increases steadily for users from the lowest grade levels to grade 12. However, after a grade of 12, the partial dependence plateaus and even slightly decreases.

This pattern suggests that users with higher grade levels are initially more likely to receive deltas, potentially reflecting greater engagement, content quality, or ability to understand counter arguments. The concave shape may also imply different behavioural dynamics between novice and more experienced writers, warranting further investigation.

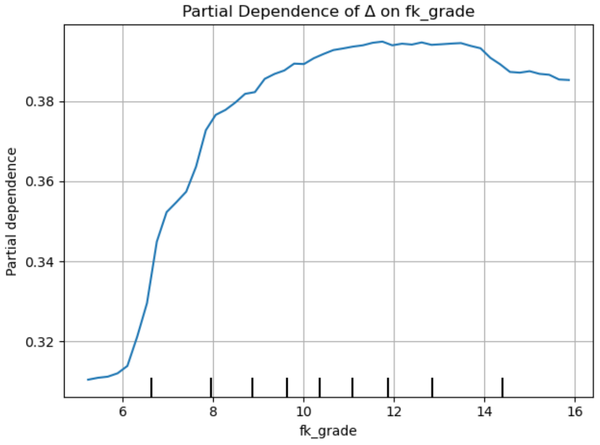


Figure 1: Partial dependence of `fk_grade` on delta prediction

5.5 Scalability Analysis

To evaluate the distributed efficiency of our pipeline, we measured processing time across different implementations of our predictive models. More specifically, we compared runtimes across distributed and non-distributed modelling setups. Due to the simplicity of base decision trees, no gains were made in the runtime of the distributed framework. However, significant improvement was made in the random forest classifier. Native modelling of a random forest classifier had a runtime of 49.78 seconds, whereas the distributed approach only took 6.89 seconds. This comparison was made on the modelling parameters with the strongest predictive power. The use of random forest classifiers in this context allows for distributed computing methods to provide the most gains in efficiency. Although these methods showed far stronger efficiency, they showed slightly weaker predictive strength in both decision tree and random forest classification, as shown in previous sections.

6 CONCLUSION & FUTURE WORK

This work presents a distributed approach to analysing persuasion dynamics in the `r/ChangeMyView` subreddit. Leveraging distributed computing techniques allowed us to process a corpus of posts and identify features associated with successful persuasion attempts. These features and associations can help us understand what drives persuasion in this setting, and how the way one approaches a forum of online discourse says about their beliefs before anyone may try and change their mind.

In the predictive setting, incorporating these features as well as the categorisation of posts in a decision tree and random forest settings showed reasonable predictive strength in classifying delta assignment by an OP on a given post, or in other words, predicting whether a user's opinion would be changed. When controlling for the imbalance in the classes, our best-performing random forest model reached a precision of .84, recall of .62 and an F1 score of .67.

The takeaway from these results is not to begin classifying all opinions in public discourse as 'changeable' or not, but instead that there is value in the analysis of how somebody approaches the forum of discussion, as well as how often others approach them. Our findings suggest that persuasion in an online setting is influenced primarily by structural factors, followed by specific linguistic strategies. Effective persuasion in this context is the culmination of many and perhaps innumerable factors, and yet information extracted from a post and its accompanying metadata are able to offer insight into human behaviour.

Future research that continues to explore the dynamics of persuasion can look to incorporate:

- Network effects and user history into the prediction model;
- More sophisticated NLP techniques such as transformer-based embeddings;
- Longitudinal analysis to track how persuasion dynamics evolve over time;
- Other subreddits, forums, or websites to either add to the corpus or compare persuasion patterns across different communities.

A challenge in studying online persuasion at scale and drawing valuable conclusions is obtaining large, high-quality datasets. CMV offers a unique circumstance of users considering themselves open to a change in their view, a characteristic hard to guarantee in online discourse. As a result, expanding this work to explore interactions beyond CMV means potentially loosening the assumptions on what we consider honest or faithful discussion. A major obstacle to our work was sourcing data from Reddit. Earlier research benefitted from Reddit's now restricted dataset via Google Big Query or from Pushshift's API. We initially pursued two alternative routes. First, through various mass scraping methods we yielded less than 1,000 posts due to API rate limits. Second, we successfully parsed a 3.8 million post on PySpark[3]from which we only reached 9,000 CMV relevant entries, significantly fewer than the publicly available data. Changes in Reddit's data policies have made complete scraping and extracting comment chains much more difficult, limiting access to rich conversational contexts that are crucial for further persuasion analysis. Finally, we also note that biases are likely to exist in the

[3] https://github.com/webis-de/webis-tldr-17-corpus

user base of Reddit, both geographically and demographically, as well as across subreddits themselves, should further research pursue them. For example, many political posts in our dataset heavily focused on Western politics, and the United States in particular. Future research could benefit from investigating the geographic distribution of posts to better understand the global representativeness of the data and to move beyond the limitations of WEIRD (Western, Educated, Industrialized, Rich, and Democratic) populations that often dominate online platforms. Broadening the study of online persuasion both in scale and diversity will require overcoming aforementioned data access constraints to find generalisable and reliable insights on online opinion change.

7 STATEMENT ON INDIVIDUAL CONTRIBUTIONS

All four group members contributed equally to the project. Tasks involving the Google Cloud Platform were collaboratively handled by all members. Two members primarily focused on data scraping and feature engineering, while the other two concentrated on developing predictive models and visualizations. All members worked equally on writing up the report. Further details on individual contributions are provided in the accompanying file `Statement_about_individual_contributions.xlsx` found in the project's Github.

8 STATEMENT ABOUT THE USE OF GENERATIVE AI

Following The London School of Economics and Political Science Generative AI guidance for taught students, we disclose that Generative AI tools were utilized during the development of this project. Specifically, OpenAI's ChatGPT was employed to assist with:

- Improve the clarity and conciseness of written explanations.
- Summarizing academic literature.
- Assist with formatting tables and structuring LaTeX outputs.
- Clarify programming syntax.
- Creating visualizations in PySpark

All AI-generated content has been critically reviewed, edited, and integrated by us to ensure it aligns with the objectives of the assignment. We have compiled and submitted all relevant chat logs detailing our interactions with ChatGPT. These logs are included in the accompanying file `AI_Chat_Logs.pdf` submitted alongside this report.

REFERENCES

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

[2] Christopher Hidey and Kathleen McKeown. 2018. Persuasive Influence Detection: The Role of Argument Sequencing. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 5173–5180.

[3] Guido W. Imbens and Donald B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

[4] Mikel Galar Isaac Triguero. 2023. *Large-Scale Data Analytics with Python and Spark: A Hands-on Guide to Implementing Machine Learning Solutions*. Addison-Wesley.

[5] Nikolay Kolyada, Khalid Al-Khatib, Michael Völske, Shahbaz Syed, and Benno Stein. 2020. *Webis ChangeMyView Corpus 2020 (Webis-CMV-20)*. https://doi.org/10.5281/zenodo.3778298

[6] Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument Strength is in the Eye of the Beholder: Audience Effects in Persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Vol. 1. 742–753.

[7] Michael Barthel, Galen Stocking, Jesse Holcomb and Amy Mitchell. 2016. *The Role of Reddit in the News Landscape*. Technical Report. Pew Research Center. https://www.pewresearch.org/wp-content/uploads/sites/8/2016/02/PJ_2016.02.25_Reddit_FINAL.pdf Accessed: 2025-05-05.

[8] Maria Obedkova. 2020. Topic Modelling with PySpark and Spark NLP. Medium blog post. https://medium.com/trustyou-engineering/topic-modelling-with-pyspark-and-spark-nlp-a99d063f1a6e

[9] Brian L. Ott and Eric Aoki. 2020. Opinion Change in 140 Characters: Testing Issue Framing, Persuasion and Credibility via Twitter and Online News Media. *Communication Studies* 71, 4 (2020), 421–437. https://doi.org/10.1080/10510974.2020.1773063

[10] Judea Pearl. 2009. *Causality* (2nd ed.). Cambridge University Press.

[11] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-Faith Online Discussions. *Proceedings of the 25th International Conference on World Wide Web* (2016), 613–624.

[12] Yu Wang, Hongxia Bai, Matt Stanton, Wen-Yen Chen, and Edward Y. Chang. 2018. PLDA: Parallel Latent Dirichlet Allocation for Large-Scale Applications. In *Proceedings of the 5th International Conference on Algorithmic Aspects in Information and Management*. 301–314.

[13] Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica. 2016. Apache Spark: A Unified Engine for Big Data Processing. *Commun. ACM* 59, 11 (2016), 56–65.

A APPENDIX

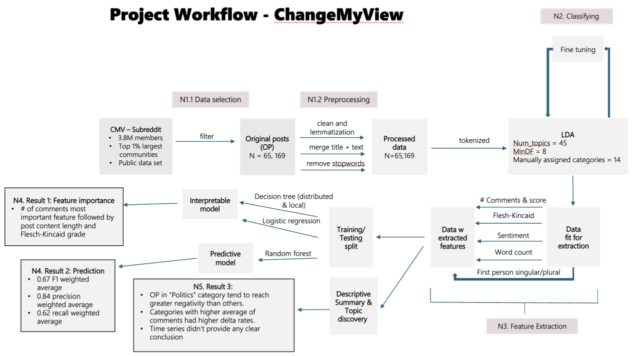

Project Workflow - ChangeMyView
N2. Classifying
Fine tuning
N1.1 Data selection
N1.2 Preprocessing
CMV – Subreddit
• 3.8M members
• Top 1% largest communities
• Public data set
filter
Original posts (OP)
N = 65, 169
clean and lemmatization
merge title + text
remove stopwords
Processed data
N=65,169
tokenized
LDA
Num_topics = 45
MinDF = 8
Manually assigned categories = 14
N4. Result 1: Feature importance
• # of comments most important feature followed by post content length and Flesch-Kincaid grade
Interpretable model
Decision tree (distributed & local)
Logistic regression
Training/ Testing split
Data w extracted features
# Comments & score
Flesh-Kincaid
Sentiment
Word count
First person singular/plural
Data fit for extraction
Predictive model
Random forest
N4. Result 2: Prediction
• 0.67 F1 weighted average
• 0.84 precision weighted average
• 0.62 recall weighted average
N5. Result 3:
• OP in "Politics" category tend to reach greater negativity than others.
• Categories with higher average of comments had higher delta rates.
• Time series didn't provide any clear conclusion
Descriptive Summary & Topic discovery
N3. Feature Extraction

Figure 2: CMV project workflow. N1 refers to notebook 1, N2 to notebook 2 and so on.

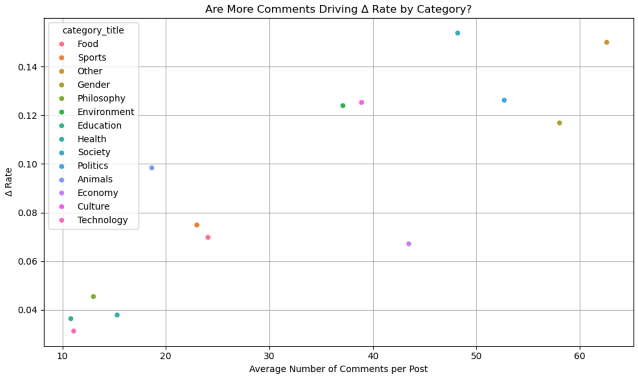
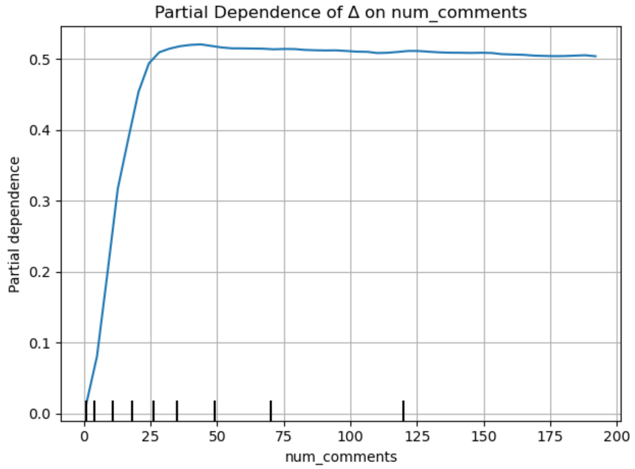


Figure 5: Number of Comments vs. Delta Success Rate, by Category



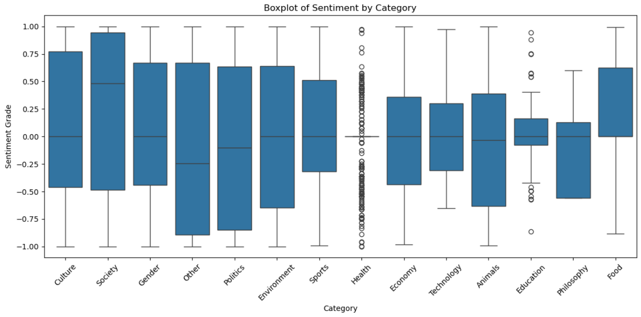


Figure 3: Boxplots of Sentiment Score by Category

Figure 6: Partial Dependence of `num_comments` on delta prediction

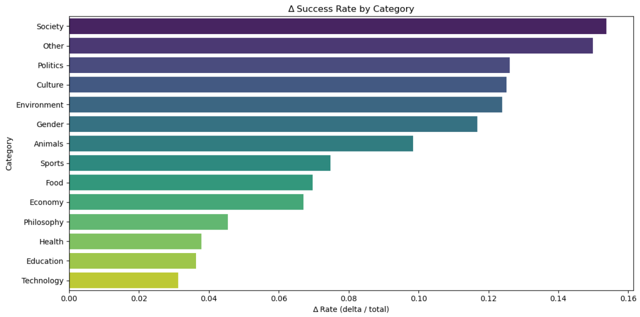


Figure 4: Delta Success Rate by Category

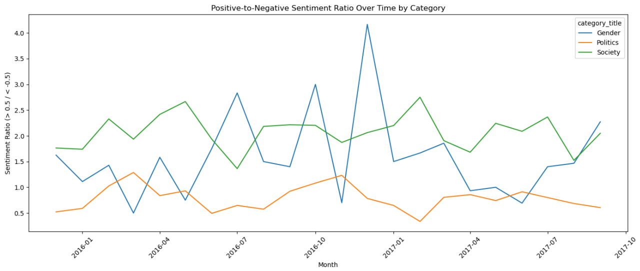


Figure 7: Number of Highly Positive or Highly Negative Comments over time by Category

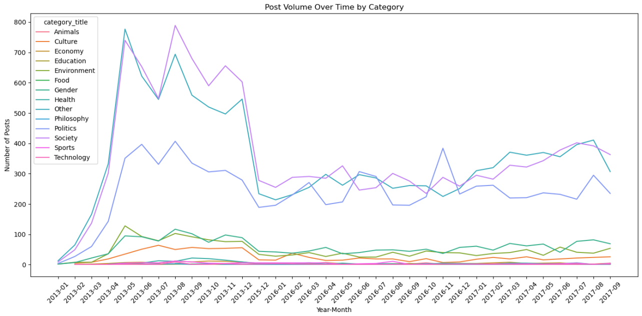


Figure 8: Number of Posts Over time, by Category

Category Title	Sum (Delta)	Avg (Num Comments)
Food	3	24.070
Sports	14	22.995
Other	3025	62.596
Gender	411	58.034
Philosophy	1	13.000
Environment	349	37.105
Education	2	10.818
Health	15	15.290
Society	3498	48.198
Politics	1655	52.692
Animals	12	18.639
Economy	20	43.473
Culture	201	38.910
Technology	2	11.094

Table 9: Summary statistics by category

Category Title	Delta Sum	Total Posts	Delta Rate
Society	3498	22749	0.1538
Other	3025	20176	0.1499
Politics	1655	13118	0.1262
Culture	201	1605	0.1252
Environment	349	2816	0.1239
Gender	411	3518	0.1168
Animals	12	122	0.0984
Sports	14	187	0.0749
Food	3	43	0.0698
Economy	20	298	0.0671
Philosophy	1	22	0.0455
Health	15	396	0.0379
Education	2	55	0.0364
Technology	2	64	0.0313

Table 10: Delta rate statistics by category